Convex Optimization and Machine Learning

Mengliu Zhao

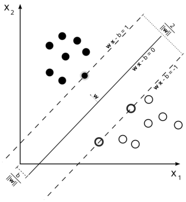
Machine Learning Reading Group
School of Computing Science
Simon Fraser University

March 12, 2014

Introduction

Formulation of binary SVM problem:

Given training data set



$$D = \{(x_i, y_i) | x_i \in R^n, y_i \in \{-1, 1\}, i = 1, 2, ..., m\} \quad (1)$$

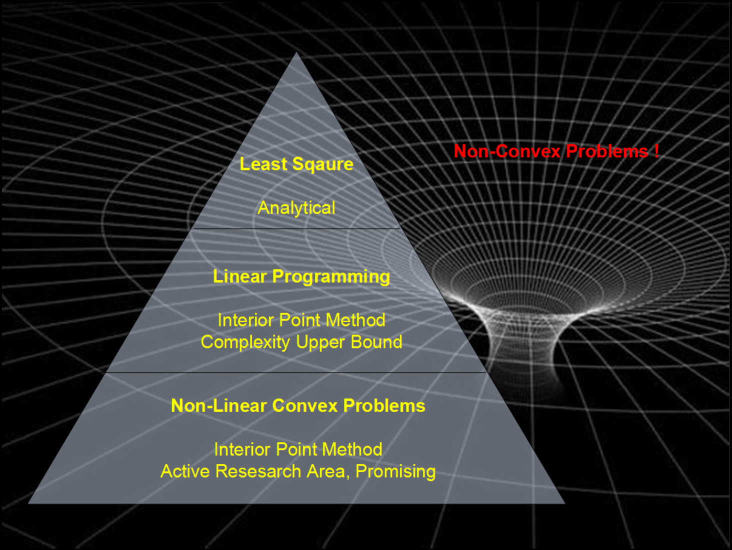
We're trying to find the maximal-margin hyperplane, which can be described by its normal vector w which satisfies (b is some offset):

$$\begin{array}{ll} \text{minimize} & ||w||_2 \\ \text{subject to} & y_i(wx_i - b) \geq 1 \quad i = 1, 2, ..., m \end{array} \quad (2)$$

Comment

We encounter a lot of constraint minimization problems in Machine Learning.

Why We Want Convex Problems?



Outline

1. Lagrange Dual Form
2. Dual Decomposition, Augmented Lagrangian and ADMM
3. SVM and Convex Optimization

Convex Optimization Problems

General form of convex optimization problem is like following:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, 2, ..., m \\ & h_j(x) = 0, \quad j = 1, 2, ..., n \end{array} \tag{3}$$

where f_0, f_i are convex functions, h_j are linear functions.

Property

The feasible set of a convex optimization problem is also convex.

In other words, convex optimization problem is solving a convex function over a convex space.

General Constraint Problem with Lagrange Duality

However, most constraint problems we optimize are not convex:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, 2, ..., m \\ & h_j(x) = 0, \quad j = 1, 2, ..., n \end{aligned} \tag{4}$$

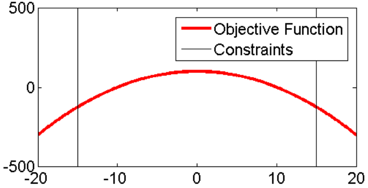
Lagrangian:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x) \tag{5}$$

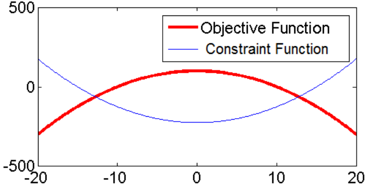
$\lambda_i (> 0), \nu_j$ are called Lagrangian multipliers or dual variables; the Lagrangian dual function is defined as:

$$g(\lambda, \mu) = \inf_x L(x, \lambda, \nu) \tag{6}$$

Geometric Explanation – Primal Problem

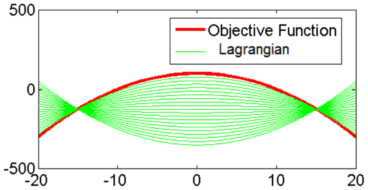


$$\begin{array}{ll} \text{minimize} & -x^2 + 15^2 \\ \text{subject to} & x^2 - 15^2 \leq 0 \end{array} \tag{7}$$



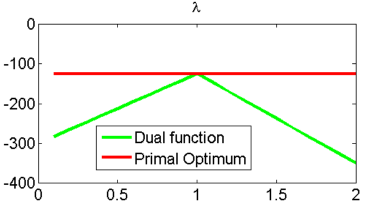
$$(x^2 - 15^2) \tag{8}$$

Geometric Explanation – Dual Problem



$$-x^2 + 15^2 + \lambda(x^2 - 15^2) \quad (9)$$

$$\lambda = .1 : .1 : 2 \quad (10)$$



$$g(\lambda) = \inf_x \{-x^2 + 15^2 + \lambda(x^2 - 15^2)\}$$

$$\lambda = .1 : .1 : 2 \quad (11)$$

Geometric Explanation – Two Observations

Observation (I)

Dual function $g(\lambda)$ is concave.

$$\begin{array}{ll} maximize & g(\lambda) \\ subject\ to & \lambda > 0 \end{array} \tag{12}$$

is a convex optimization problem.

Observation (II)

Let p^* be the optimal value of the primal problem, then

$$g(\lambda) \leq p*, \forall\lambda \tag{13}$$

Economic Explanation

Company production cost f_0, with certain limits f_i below a_i (rules, resources):

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) - a_i \leq 0, \quad i = 1, 2, ..., m \end{aligned} \tag{14}$$

However, if , the company can pay a fund rate of $\lambda_i > 0$ to violate certain rules, which adds back to the total cost:

$$g(\lambda) = \inf_{\lambda} \{ f_0(x) + \sum_i \lambda_i (f_i - a_i) \} \tag{15}$$

In this case, the optimal value d^* for the company is the cost under the least favorable set of prices $\lambda \longrightarrow \max g(\lambda)$.

Strong & Weak Duality

How well does the dual problem approximate the original problem?

1. Weak Duality: optimal duality gap is always non-negative.

$$p^* - d^* \geq 0 \tag{16}$$

2. Strong Duality: duality gap is zero.

$$p^* = d^* \tag{17}$$

   Q: When does strong duality hold?

Theorem (Slater's Theorem)

D is feasible set. Assume the primal problem is convex:

$$\begin{array}{lll} \mathit{minimize} & f_0(x) & \\ \mathit{subject\ to} & f_i(x) \leq 0, & i = 1, 2, ..., m \\ & h_j(x) = 0, & j = 1, 2, ..., n \end{array} \tag{18}$$

If $\exists x \in \operatorname{relint} D$, and $f_i(x) < 0, i = 0, 1, ..., m$, then strong duality holds.

KKT Condition

For constrained problem:

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, 2, ..., m \\ & h_j(x) = 0, & j = 1, 2, ..., n \end{array} \tag{19}$$

If x^* is the primal minimum, then it satisfies the following necessary condition:

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x^*) + \sum_{j=1}^{n} \nu_j \nabla h_j(x^*) = 0 \tag{20}$$

Outline

Dual Form, Then What?

Once we get the dual problem, it's easy to solve, e.g., by gradient approach (dual ascent).

Property

If $g(\lambda)$ is a convex (concave) function, then $\nabla f(\lambda^*) = 0$ iff λ^* is the global minimizer (maximizer).

Comment

A lot of conditions need to be satisfied for a stable gradient method.

Dual Ascent for Solving Dual Problem

Let's look at a simplified version of the constrained problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array} \tag{21}$$

Its dual form:

$$\max g(\lambda) = \max_{\lambda}\{\min_{x} L(x, \lambda)\} \tag{22}$$

$$L(\lambda, x) = f(x) + \lambda(Ax - b) \tag{23}$$

Update x, λ at each iteration:

$$x^{k+1} = \min_{x} L(x, \lambda^k) \tag{24}$$

$$\lambda^{k+1} = \lambda^k + \alpha^{k+1} \nabla g(x^{k+1}, \lambda^k) \tag{25}$$

Question

What if we have a much more complex situation?

Dual Decomposition

Suppose the problem is of high dimension, $\hat{x} = (x, z)$, and $f(\hat{x})$ is separable:

$$
\begin{aligned}
f(\hat{x}) &= f_1(x) + f_2(z) && (26)\\
A\hat{x} - b &= (A_1 x - b_1) + (A_2 z - b_2) && (27)
\end{aligned}
$$

Then we can do dual ascent on each dimension separately:

$$
\begin{aligned}
L_1(x) &= f_1(x) + \lambda_1(A_1 x - b_1) && (28)\\
L_2(x) &= f_2(x) + \lambda_2(A_2 x - b_2) && (29)\\
x^{k+1} &= \min_x L_1(x, z^k, \lambda^k) && (30)\\
z^{k+1} &= \min_z L_2(x^{k+1}, z, \lambda^k) && (31)\\
\lambda^{k+1} &= \lambda^k + \alpha^{k+1} \nabla g(x^{k+1}, z^{k+1}, \lambda^k) && (32)
\end{aligned}
$$

Comment

1. *Simple dual ascent is usually slow;*

Alternative to Dual Ascent – Augmented Lagrangian

Primal problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array} \tag{33}$$

Dual problem:

$$L(x, \lambda, \theta) = f(x) + \lambda^T (Ax - b) + \frac{\theta}{2} ||Ax - b||_2^2 \tag{34}$$

Update by method of multipliers (fixed step):

$$x^{k+1} := \min_x L(x, \lambda^k, \theta) \tag{35}$$

$$\lambda^{k+1} := \lambda^k + \theta(Ax^{k+1} - b) \tag{36}$$

Method of Multipliers

Comparing to dual ascent:

1. *Good news*: convergence under more relaxed conditions;
2. *Bad news*: dual decomposition no longer works (now we have quadratic terms)!

Comment

ADMM can help!

ADMM

Alternating Direction Method of Multipliers

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = b \end{array} \tag{37}$$

Its Lagrangian is:

$$L_\theta(x, \lambda, z) = f(x) + g(z) + \lambda^T(Ax + Bz - b) + \frac{\theta}{2}||Ax + Bz - b||_2^2 \tag{38}$$

ADMM scheme:

$$\begin{aligned} x^{k+1} &:= \min_x L_\theta(x, z^k, \lambda^k) \\ z^{k+1} &:= \min_z L_\theta(x^{k+1}, z, y^k) \\ \lambda^{k+1} &:= \lambda^k + \theta(Ax^{k+1} + Bz^{k+1} - b) \end{aligned} \tag{39}$$

A Closer Look at ADMM

Comment

We need more convincing evidence that the scheme will work!

The thing unnatual here is the new variable z. We'll check the KKT condition with the constraint problem above:

$$\nabla g(z) + B^T\lambda = 0 \tag{40}$$

We'll check if this could be satisfied by the ADMM scheme. Since z^{k+1} minimized $L_\theta(x^{k+1}, z, \lambda^k)$, then

$$\begin{aligned} 0 &= \nabla g(z^{k+1} + B^T\lambda^k + \theta B^T(Ax^{k+1} + Bz^{k+1} - b)) && (41) \\ &= \nabla g(z^{k+1} + B^T\lambda^k && (42) \end{aligned}$$

Which means the KKT condition is satisfied.

Outline

Dual Form of SVM

Now let's come back to the constrained version of SVM model:

$$\begin{array}{ll} \text{minimize} & ||w||_2 \\ \text{subject to} & y_i(wx_i - b) \geq 1 \quad i = 1, 2, ..., m \end{array} \tag{43}$$

It's easy to convert it to Lagrangian dual form as following:

$$\max_{\lambda}\{\min_{w,b}\{||w||_2^2 + \sum \lambda_i[1 - y_i(wx_i - b)]\}\} \tag{44}$$

Comment

The formulation is too complex! We can do further to simplify it!

Dual Form of SVM

Check KKT condition, taking 1-order derivative of w and b on Lagrangian function $||w||_2^2 + \sum \lambda_i [1 - y_i(wx_i - b)]$:

$$
\begin{aligned}
w &= \sum_i \lambda_i y_i x_i && (45)\\
0 &= \sum \lambda_i y_i && (46)
\end{aligned}
$$

Replace them back in (44), we have:

$$
\begin{aligned}
& \max_\lambda g(\lambda) &=& \max_\lambda \{\sum \lambda_i - \tfrac{1}{2} \sum y_i y_j \lambda_i \lambda_j (x_i)^T x_j\} \\
s.t. \quad & \lambda_i &\geq& 0, i = 1, 2, ..., m \qquad (47)\\
& \sum \lambda_i y_i &=& 0
\end{aligned}
$$

Summary

1. Lagrangian Duality, KKT condition
2. Dual Decomposition, Augmented Lagrangian, ADMM
3. Example using Lagrangian Duality on SVM

Thank You!